chinaXiv:202310.01148v1

mongodb 分片集群方案设计和部署

杨宏章
（新华通讯社通信技术局 数据技术部，北京 100803）

摘　要：mongodb 已经是比较成熟的 NoSQL 数据库产品，使用也比较普遍。搭建 mongodb 分片集群，mongodb 企业版提供了 OpsManager 等辅助工具，实现了完整的自动化的部署、升级、监控、备份及恢复方案。相比之下，使用 mongodb 社区版来搭建分片集群需要一定的经验积累，本文整理了 mongodb 副本集的规则和特性，在实验环境中经过验证，力求更贴近实际，更具可读性和可操作性，并就搭建 mongodb 分片集群进行探讨，针对需要解决的问题、考虑的因素，设计出相应的方案，并对部署实践给出建议。

关键词：mongodb；社区版；分片集群；副本集；分片；投票节点；容错机制　　**中图分类号：**TP393　　**文献标识码：**A
文章编号：1671−0134（2021）03−111−03　　**DOI：**10.19483/j.cnki.11−4653/n.2021.03.031
本文著录格式：杨宏章 .mongodb 分片集群方案设计和部署 [J]. 中国传媒科技，2021（03）：111−113.

1. 软件版本

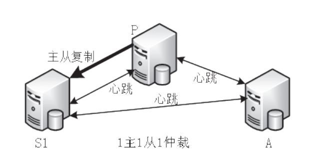
64 位操作系统 CentOS6.5+；

mongodb3.6 社区版；

2. 副本集规则和特性

2.1　副本集的成员角色、数量

副本集由 1 个主节点 P（Primary）、若干个从节点 S（Secondary）、仲裁节点 A（Arbiter，根据实际情况设定）构成。每个副本集的节点总数不超过 50，最少有 3 个节点（分片集群排除单节点、主从双节点的情况），如图 1 所示的 2 种情况：1 主 2 从，1 主 1 从 1 仲裁，单箭头表示主从复制，双箭头表示心跳。



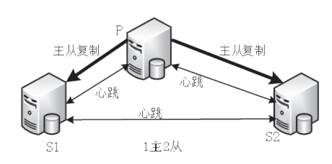


图 1　三个节点的副本集结构

2.2　副本集选举的大多数

选举主节点，需得到大多数成员的支持，这里的大多数，是具有投票权的节点的半数以上， 不一定是副本集节点总数的大多数。

副本集节点总数	默认投票节点总数	大多数	容错数
3	3	≥ 2	1
4	4	≥ 3	1
5	5	≥ 3	2
6	6	≥ 4	2
7	7	≥ 4	3
8	7	≥ 4	3
9	7	≥ 4	3

图 2　副本集选举的大多数

2.3　有投票权的节点总数

一个副本集中，有投票权的节点总数不超过 7 个；若副本集的节点总数不超过 7，默认每个节点都有投票权；若副本集节点总数多于 7 个，超出 7 个之外的差额数量的节点，必须设成没有投票权。比如：副本集节点总数为 9，必须设定 2 个节点没有投票权。即：若投票节点数量少于副本集节点总数，需要设定差额数量的节点没有投票权。

图 2 中的容错数，指在能正常选举出主节点的情况下，投票节点失效的最大数量。只要失效的投票节点不超过容错数，就能正常选举出主节点。

副本集节点总数、有投票权的节点总数，一般都设定为奇数，但不是强制性的。关于节点数量设成奇数还是偶数，有一种观点认为：如果设成偶数，选举的时候可能出现 2 个节点得票数一样多的情况，从而选不出主节点。我们假设这种情况成立，如：节点总数为 4，可用节点数为 4，可能选不出主节点；再考虑一下节点总数为 5，投票节点数量为 5，对 1 个节点停机维护，剩下 4 个节点可用，也可能选不出主节点。这显然是个悖论，和 mongodb 高可用架构的特性不符。经实验验证，设定为偶数个，只要选举超过半数，同样能选出主节点。但是，从图 2 中可以看出，奇数个投票节点，再增加一个成为偶数个，不能提高容错数，反而降低了副本集的稳定性。如：3 个节点，2 个可用（67% 的节点可用）就能达到大多数； 4 个节点需要 3 个可用（75% 的节点可用）才达到大多数。

2.4　节点的投票权

可通过参数 votes 来设定，votes 默认值为 1，表示可投 1 票（值不能大于 1，不允许一个节点可以投多票的情况发生），votes 值为 0，则无投票权（不考虑投否决票；其优先级 priority 值也必须为 0）。

2.5　节点的优先级

通过参数 priority 来设定， priority 取值范围是从 0 到

ChinaXiv合作期刊

1000之间的浮点数，[1] 默认值为1，数值越大，优先级越高，越容易被选举成主节点。如果希望某个成员能被选为主节点，可调高其优先级。

2.6 仲裁节点

仲裁节点只参与选举，不存放数据（不会被选举成主节点），设置成仲裁节点后，不能更改成从节点，只能从副本集中移除后，重新初始化再加入副本集。仲裁节点可依据实际需要设置，额外的仲裁节点，既不能提高数据安全性，还会影响选举效率。

2.7 隐藏节点

对客户端是不可见的，客户端不会向隐藏节点发送请求。[2] 设置成隐藏节点，可避免客户端的请求，虽然隐藏节点不会成为主节点，但在选举过程中可以正常投票。参数设置：priority：0， hidden：true。

2.8 延迟节点

延迟节点必须是隐藏节点，选举过程中可以正常投票，不会收到客户端的请求。延迟节点的数据会比主节点延迟指定时间，不会成为主节点，参数设置，如：priority：0，hidden：true，slaveDelay：259200（单位：秒，即延迟 3 天）。

2.9 投票节点没有达到大多数的情况

副本集中投票的节点若没有达到（有投票权的节点总数）半数以上，不能选举出新的主节点，并且当前主节点会自动降级成为从节点。

3.副本集设计

3.1 容错机制

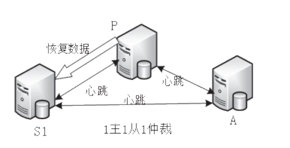
为应对各类突发事件，如：节点数据损坏、物理机宕机、网络连接中断，等等，在副本集的部分节点不可用的情况下，副本集要正常提供服务，对容错数是有要求的。

容错数为 n，表示：n 个投票节点失效的情况下，副本集可以正常选举出主节点。副本集中节点数量最少为 2n+1，如：容错数为 2，副本集的节点总数最少为 5。

3.2 单个数据节点的故障恢复

数据节点的故障恢复，依据损坏情况，用不同方案。如果只是数据损坏，硬件等其它条件都正常，只需恢复数据。如果硬件类故障，需另外准备好一个新的节点后，再恢复数据。

图 3 是向 S1 节点恢复数据，1 主 1 从 1 仲裁的结构，数据节点只有 2 个，数据源只能选择主节点 P，大数据量的复制会对服务器和应用程序造成显著影响。右图是 1 主 2 从结构，数据节点多于 2 个，可以避免选择主节点作为数据源，因为可以选择从节点 S2 作为数据源。



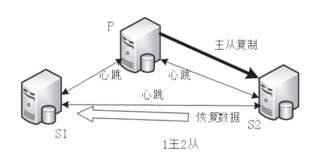


图 3 节点恢复数据

3.3 仲裁节点的使用

仲裁节点的使用，可以保证副本集的投票节点数量为奇数，提高副本集的稳定性，同时，因为仲裁节点不存放数据，也减少了数据冗余。此外，在异地、跨机房的部署中，仲裁节点有特殊的作用。

3.4 防止数据误操作：延迟节点的使用

恢复误操作的数据，需要用到延迟的备份节点 D（Delay node）。因为主、从复制延迟了一段时间，这段时间内对数据的操作，延迟节点还没有执行。延迟多久，依据防止数据误操作的需求，人为设定的，如：延迟 3 天，可恢复 3 天内的误操作数据；延迟 7 天，可恢复 7 天内的误操作数据。

同时，依据业务量，数据规模，在节点配置文件中设置好参数 oplogSizeMB 的大小，默认大小是可用磁盘空间的 5%，保证 oplog 日志保留时间比节点的延迟时间要长。

3.5 指定主节点、从节点

若硬件资源有区别，建议把性能最好的设定成主节点，性能一般的设定成从节点，性能稍差的设定成仲裁节点。

在跨机房部署的情况下，一般希望把主节点部署在主机房。

通过设定优先级指定主节点还有一个好处，就是可以避免出现一种极端情况：多个分片，节点关闭、启动引起选举，若使用默认优先级，各分片的主节点被选出来后，所有的主节点都在同一主机上，集群的资源分配是不合理的，通过指定主节点，就完全解决了这个问题，主节点分布在不同主机上，也能实现业务的分布式处理。

3.6 votes：0，priority：0 节点的使用

设置参数 votes：0，priority：0 的节点，不参与选举，不会接收到客户端的请求，只作为数据备份节点。

4.集群部署

4.1 副本集节点跨机房、异地部署

为满足异地、或不同机房的灾备要求，需采用多中心架构。如图 4：

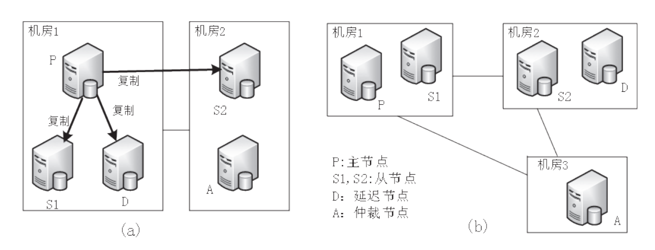


图 4 副本集节点跨机房部署

机房 1 是主机房，通过优先级把主节点部署在主机房。（a）图：可通过副本集参数调整，把主节点切换到机房 2；在两个机房之间的通信中断的情况下，主机房依旧能正常选举；但是，如果机房 1 出现整个机房故障，就不能保障副本集可用，因为机房 2 达不到半数以上，不能选举

chinaXiv:202310.01148v1

chinaXiv:202310.01148v1

出新的主节点。为解决这个问题，（b）图使用了机房 3，用于部署仲裁节点，同时机房 1 和机房 2 节点数量相等，无论机房 1 或机房 2 出现整个机房故障，仲裁节点都能和另一机房节点构成大多数，保证副本集可用。

隐藏节点 D、仲裁节点 A 虽然都不会成为主节点，但都有投票权。

4.2 副本集节点的分布式部署

同一个副本集的节点若部署在同一主机上，会出现主机宕机导致整个副本集不可用的情况。为避免这种情况，副本集的节点，要分别部署在不同的主机上。

4.3 单机单节点 / 单机多节点部署

单机单节点，即：一个主机上只部署一个节点。单机多节点，即：一个主机上部署多个节点（这些节点分属不同的副本集）。分片集群一般有多个分片，也就有多个副本集，若几个节点分属于不同的副本集，是可以部署到同一个主机上的。一个主机上部署多少个节点，需要考虑硬件资源，主要是 CPU 和内存资源。如图 5，Server1 部署了 1 个 config 节点，1 个 mongos 节点，分片 shard1 的主节点，shard2 和 shard3 的从节点。

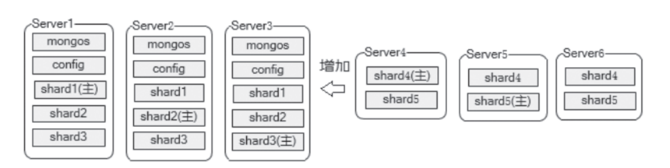


图 5 集群扩展

4.4 分片的数量以及集群的扩展

使用 mongodb 分片集群，数据规模一定很大，否则使用副本集的方案就够用了。易扩展性是 mongodb 最显著的优点之一。所以，一方面，因为易扩展性，分片数量可以设置少一些，比如 2 个分片，以后有需要可以随时增加分片；另一方面，数据均衡器已经把历史数据大致均匀地存放到不同分片中，新增加分片，会引发向新的分片中迁移数据，既费时间，又耗系统资源。因此，增加分片的操作不能太频繁。只有对承载的业务量、数据量进行充分调研和预估，设计的方案才有前瞻性，在一段时期内不需要增加分片。

对集群进行扩展，如图 5，Server1，Server2，Server3 组成了 3 分片 1 主 2 从 3 副本的集群，扩展时增加了 Server4，Server5，Server6，增加了 2 个分片 shard5 和 shard6，也是 1 主 2 从结构。

4.5 同一组服务器部署多个集群

生产系统不建议这么部署。如果硬件性能确实够用，用于实验目的，也可以尝试。在每一个节点的配置文件中，都有一个参数 security.keyFile，用于指定验证文件，同一个 mongodb 集群的节点使用的验证文件相同。

4.6 config（配置）节点的部署

mongodb3.2 版，配置节点支持 2 种方式（任选一种）：

SCCC Config Server：多个 config 节点以镜像的方式；

CSRS Config Server：多个 config 节点，组成副本集；

3.2 版之前，只支持 SCCC 方式，3.2 版之后，只支持 CSRS 方式。从 3.2 版升级到 3.4 版，首先要把配置节点设置成副本集。

参考文献

[1]MongoDB Manual[OL] https：//docs.mongodb.com/

[2] Kristina Cbodorow 著，邓强、王明辉译，MongoDB 权威指南（第二版）[M]. 北京：人民邮电出版社，2014（1）.

作者简介：杨宏章（1982-），男，湖北黄冈，新华通讯社通信技术局工程师。

（责任编辑：张晓婧）

（上接第110页）

络资源的虚拟化，使得资源利用率最大化，取得了很好的经济效益。

4. 结语与展望

随着国家政府对政务上云、工业企业上云和金融系统上云等一系列政策的推动和投资，相信公有云服务将得到大力发展，就正如基于对象的云存储服务不断发展，其海量、安全、高可靠、低成本的数据存储能力将使得新闻服务平台实现数据的一次搬迁，全球通用，将大大提高亚太站点向北美、欧洲站点数据同步的效率。同时随着公有云 PaaS 及 SaaS 层服务不断地推陈出新，未来一定会促进新闻服务平台的巨大变革。

参考文献

[1] 孟雨 . 中国公有云服务市场达 84 亿美元 [J]. 计算机与网络，2020，（21）：12.

作者简介：杨旗（1988-），男，湖北省荆州市，工程师。

（责任编辑：张晓婧）